Environmental Health Perspectives
Vol. 87, pp. 143–147, 1990

A Comparative Study of Two Statistical Models for the Analysis of Binary Data from Longitudinal Studies

by Hideki Origasa* and James D. Knoke†

This study extensively compares two statistical models for the analysis of binary data from longitudinal studies. The first model was proposed by Zeger, Liang, and Self, which was abbreviated as ZLS model and another model was proposed by Origasa. The comparison focuses on both analytical and statistical viewpoints. The first discusses a type of the models and the second evaluates the effect from model misspecification by simulation, assuming that the ZLS model is true.

Introduction

A study in health sciences frequently uses a design involving a time factor. Data are collected at multiple occasions with respect to each subject. They may be referred to as longitudinal data or repeated measures. Such data can be produced by both retrospective and prospective studies. Survival data are usually excluded from them because they cannot involve recurrent events (*1*). Thus, a longitudinal study may be defined as one in which data are collected on several occasions, regardless of the direction and type of study.

The longitudinal study provides several advantages over the cross-sectional study. For example, it increases the precision of treatment contrasts by eliminating within-individual variation and enables us to examine the individual's changing response pattern over time.

Time series technique may be a solution for analysis of such dependent observations. However, it is only effective for studies with a large number of occasions. A study in health sciences often involves relatively small number of occasions, say two to six. In most of the clinical trials conducted by Japanese pharmaceutical companies, data are collected on a few occasions after randomization. Also, from the viewpoint of modeling, the data need to include some covariates such as baseline risk factors into the model.

Literature Review

Three approaches are possible for analyzing binary longitudinal data. The first is modeling for marginal probabilities; the second is modeling for transition (or conditional) probabilities; and the last is nonparametric, i.e., not a model-based approach.

With respect to the first type of modeling, GSK (Grizzle, Starmer, and Koch) linear model (*2*) is fairly general. It can be applied to longitudinal data (*3*). Suppose there are T occasions with binary responses. Then, there are 2^T profiles that each individual corresponds to. On can express any function generated from the vector of profiles. Another function is shown by Liang and Zeger (*4*) which uses the generalized linear model (*5*). The within-individual covariance matrix is included in the model. This model allows us to deal with a mixture of discrete and continuous variables.

Modeling for transition probabilities has been proposed by many authors. Again, the GSK approach is applicable for them. The Markov chain model can also be applied. Muentz and Rubinstein (*6*) has shown a logistic expression for those. The last two, i.e., ZLS (*7*) model and Markov logistic regression model (*8*) will be described in another section.

By rearranging data into T consecutive 2 × 2 tables (Table 1), several authors have proposed different statistics to test for treatment effect (*8–10*). An underlying model is T-fold product binomial (*11*) with Markov property.

Table 1. Rearranged 2 × 2 consecutive tables from longitudinal binary data.

	Time 1			Time T		
Group	Yes	No	Total	Yes	No	Total
Drug	a_1	b_1	$N/2$	a_T	b_T	$N/2$
Control	c_1	d_1	$N/2$	c_T	d_T	$N/2$

*Eisai Company, Ltd., Tokyo 112, Japan.
†George Washington University, Rockville, MD 20852.
Address reprint requests to H. Origasa, Eisai Company, Ltd., Tokyo 112, Japan.

ZLS Model

The ZLS model is formulated as the following two stages. The first one is expressed as:

$$\text{logit}\ (p_{i1}) = \log\{p_{i1}/(1 - p_{i1})\} = \mathbf{Z}_i'\delta,$$

at the initial occasion $\mathbf{Z}_i$ is a $q \times 1$ vector of time-independent covariates and δ is a vector of associated parameters. The second stage expresses the stationary first-order autoregressive, that is,

$$p_{it} = p_{i1} + \rho\,(y_{i,t-1} - p_{i1}),\ t \geq 2$$

where ρ is the autocorrelation coefficient.

Time course is simply determined by the most recent outcome, autocorrelation coefficient, and the initial probability so that no covariates are related to changing probabilities of having a symptom over time. Statistical inference can be performed using the likelihood, that is,

$$L = \prod_{i=1}^{N}\left[p_{i1}^{y_{i1}}(1 - p_{i1})^{1-y_{i1}}\left\{\prod_{t=2}^{T}p_{it}^{y_{it}}(1 - p_{it})^{1-y_{it}}\right\}\right]$$

which is called the unconditional likelihood because it summarizes the entire data.

Markov Logistic Regression Model (MLRM)

This model comes from a small modification of the ordinary logistic regression model to incorporate the covariate of previous outcome. It approximately corresponds to a covariance structure of the first-order autoregressive process. It allows us to use data much more efficiently than the multivariate approach, such as the one proposed by Grizzle and Allen (*12*). The principle is that adjacent nonmissing pairs can be used. The model is expressed as:

$$\text{logit}\ (p_{it}) = \alpha + \beta\, y_{i,t-1} + \mathbf{X}_{it}'\,\gamma + \mathbf{Z}_i'\delta$$

where the p_{it} is the conditional probability of having a response at time t ($t = 1, \ldots, T$) for the ith individual, given the past observation ($y_{i,t-1}$) and the covariates ($\mathbf{X}_{it}$, $\mathbf{Z}_i$).

Although the principle models transition probabilities, it exactly corresponds to the modeling for marginal probabilities denoting $p_{ij}(i,j = 0 \text{ or } 1)$ to be transition probabilities and $\pi_{ij}(i,j = 0 \text{ or } 1)$ to be marginal probabilities. Define

$$p_{01} = Pr\{Y_2 = 1 \mid Y_1 = 0\},\ p_{00} = 1 - p_{01},$$

$$p_{11} = Pr\{Y_2 = 1 \mid Y_1 = 1\},\ p_{10} = 1 - p_{11},$$

$$\pi_{01} = [1 - Pr\{Y_1 = 1\}] \times p_{01},$$

$$\pi_{00} = [1 - Pr\{Y_1 = 1\}] \times p_{00},$$

$$\pi_{11} = Pr\{Y_1 = 1\} \times p_{11},$$

$$\pi_{10} = Pr\{Y_1 = 1\} \times p_{10}$$

in which the first and second occasions are concerned. Consider the MLRM without covariates, that is,

$$\text{logit}\ (p_{it}) = \alpha + \beta\, y_{i,t-1}$$

In other expressions,

$$\log\{p_{01}/p_{00}\} = \alpha,\ \log\{p_{11}/p_{10}\} = \alpha + \beta$$

depending on the realization of previous outcome. From a simple algebra, we obtain

$$\log\{\pi_{01}/\pi_{00}\} = \alpha,\ \log\{\pi_{11}/\pi_{10}\} = \alpha + \beta$$

which has an equivalent form to the above.

Comparisons

Suppose that there are no time-dependent covariates. The MLRM turns out to be:

$$\text{logit}\ (p_{it}) = \beta\, y_{i,t-1} + \mathbf{Z}_i'\delta$$

or explicitly

$$p_{it} = \exp(\beta\, y_{i,t-1} + \mathbf{Z}_i'\delta)/\{1 + \exp(\beta\, y_{i,t-1} + \mathbf{Z}_i'\delta)\}$$

The ZLS model is, on the other hand, expressed as:

$$p_{i1} = \exp\,(\mathbf{Z}_i'\delta^*)/\{1 + \exp\,(\mathbf{Z}_i'\delta^*)\}$$

and

$$p_{it} = p_{i1} + \rho\,(y_{i,t-1} - p_{i1}),\ t \geq 2,$$

where the parameter δ^* is generally different from δ. Although the transition of responses is only varied by a constant autocorrelation parameter (ρ) for the ZLS model, it is a complex expression for the MLRM as:

$$\rho = [\exp(\beta y_{i,t-1} + \mathbf{Z}_i'\delta)/\{1 + \exp(\beta y_{i,t-1} + \mathbf{Z}_i'\delta)\} - p_{i1}]/[y_{i,t-1} - p_{i1}]\,.$$

A plausible expression for the relative risk from a previous outcome might be different between two models. It might be useful for the ZLS model to express it as an additive form, so that

$$R_{\text{ZLS}} = Pr\,\{y_{it} = 1 \mid y_{i,t-1} = 1\} - Pr\{y_{it} = 1 \mid y_{i,t-1} = 0\} = \rho$$

If a past observation is unrelated to the present one, then the relative risk should be zero which corresponds to $\rho = 0$. A relative risk for the MLRM might be usefully expressed as a multiplicative form as:

$$R_{\text{MLRM}} = \frac{Pr\,\{y_{it} = 1 \mid y_{i,t-1} = 1\}}{Pr\{y_{it} = 1 \mid y_{i,t-1} = 0\}} = \frac{e^{\beta} + \exp(\mathbf{Z}_i'\delta + \beta)}{1 + \exp(\mathbf{Z}_i'\delta + \beta)}$$

The null value of relative risk is 1 when $\beta = 0$, which means there is no effect from the previous outcome.

Simulation Study

The purpose of conducting a simulation study is to evaluate the robustness of the MLRM from the view-

Table 2. Effects of the model misspecification on the estimate, variance, and goodness of fit (a = 0, ρ = 0.0, c = 0.0).

		Empirical estimate			Empirical variance			
N	T	a	ρ	c	a	ρ	c	Maximum log-likelihood
ZLS model								
30	3	0.002	− 0.046	− 0.010	0.056	0.030	0.062	− 39.8
30	6	0.009	− 0.015	− 0.010	0.024	0.008	0.025	− 102.5
60	3	0.008	− 0.013	− 0.011	0.024	0.010	0.026	− 81.8
60	6	0.002	− 0.007	− 0.003	0.011	0.003	0.011	− 206.7
		α	β	γ	α	β	γ	
MLRM								
30	3	0.050	− 0.085	0.006	0.052	0.116	0.028	− 40.0
30	6	0.030	− 0.050	− 0.009	0.065	0.142	0.036	− 102.5
60	3	0.029	− 0.031	− 0.009	0.154	0.320	0.083	− 81.7
60	6	0.016	− 0.028	− 0.004	0.027	0.053	0.013	− 206.5

Table 3. Effects of the model misspecification on the estimate, variance, and goodness of fit (a = 0, ρ = 0.0, c = 0.3).

		Empirical estimate			Empirical variance			
N	T	a	ρ	c	a	ρ	c	Maximum log-likelihood
ZLS model								
30	3	0.009	− 0.034	0.304	0.050	0.023	0.057	− 39.4
30	6	0.017	− 0.017	0.296	0.025	0.009	0.024	− 100.8
60	3	0.015	− 0.019	0.298	0.028	0.013	0.026	− 80.3
60	6	0.006	− 0.007	0.299	0.013	0.004	0.011	− 203.3
		α	β	γ	α	β	γ	
MLRM								
30	3	0.054	− 0.076	0.330	0.148	0.316	0.088	− 39.4
30	6	0.041	− 0.058	0.301	0.057	0.128	0.028	− 100.9
60	3	0.044	− 0.043	0.303	0.069	0.152	0.034	− 80.4
60	6	0.020	− 0.029	0.301	0.029	0.058	0.013	− 203.2

Table 4. Effects of the model misspecification on the estimate, variance, and goodness of fit (a = 0, ρ = 0.2, c = 0.0).

		Empirical estimate			Empirical variance			
N	T	a	ρ	c	a	ρ	c	Maximum log-likelihood
ZLS model								
30	3	0.007	0.175	− 0.007	0.061	0.019	0.060	− 39.2
30	6	0.019	0.187	− 0.010	0.031	0.007	0.034	− 99.6
60	3	0.011	0.189	− 0.010	0.029	0.009	0.029	− 79.6
60	6	0.006	0.194	− 0.005	0.014	0.003	0.016	− 200.7
		α	β	γ	α	β	γ	
MLRM								
30	3	− 0.384	0.795	0.009	0.161	0.332	0.083	− 38.8
30	6	− 0.377	0.776	− 0.009	0.058	0.128	0.030	− 99.4
60	3	− 0.386	0.801	− 0.009	0.072	0.150	0.038	− 79.2
60	6	− 0.390	0.793	− 0.006	0.028	0.060	0.014	− 200.5

point of model misspecification. The focus is on the degree of similarity between two models with respect to the goodness of fit, conservativeness, and power of the likelihood ratio test, provided that ZLS model is true. The final maximum log-likelihood is used for the index of goodness of fit since the likelihood based inference is used for both models and they have an equal number of parameters. The degrees of conservativeness and power of the likelihood ratio tests are compared between two models under four different tails (i.e., 1, 5, 10, and 20%).

The maximum likelihood estimation is performed using the quasi-Newton method (*13*), and the binary observations are created by the acceptance-rejection method based on the uniformly distributed random number generator, the number of replications is 500. The number of sample sizes to be considered is 4, depending on both the number of subjects and the number of occasions.

The true ZLS model is:

$$\text{logit}\,(p_{i1}) = a + c\,\text{TRT}_i,$$
$$p_{it} = p_{i1} + \rho\,(y_{i,t-1} - p_{i1}),\ t \geq 2$$

where TRT_i has a value of 1 for the control and -1 for

Table 5. Effects of the model misspecification on the estimate, variance, and goodness of fit (a = 0, ρ = 0.2, c = 0.3).

		Empirical estimate			Empirical variance			
N	T	a	ρ	c	a	ρ	c	Maximum log-likelihood
ZLS model								
30	3	0.015	0.167	0.299	0.060	0.018	0.063	− 38.6
30	6	0.018	0.182	0.289	0.033	0.007	0.034	− 98.3
60	3	0.014	0.186	0.287	0.029	0.008	0.031	− 78.6
60	6	0.013	0.191	0.297	0.014	0.003	0.016	− 197.7
		α	β	γ	α	β	γ	
MLRM								
30	3	− 0.369	0.772	0.191	0.174	0.328	0.095	− 38.2
30	6	− 0.370	0.765	0.245	0.058	0.123	0.031	− 98.1
60	3	− 0.378	0.799	0.242	0.073	0.152	0.039	− 78.1
60	6	− 0.385	0.792	0.247	0.028	0.058	0.014	− 197.5

Table 6. Effects of the model misspecification on the significance results (a = 0, ρ = 0.0, c = 0.0).

		Type I error							
		Empirical LR power of ρ				Empirical LR power of β			
N	T	1%	5%	10%	20%	1%	5%	10%	20%
Autocorrelation effect									
30	3	8.4	11.2	14.8	24.0	1.4	5.4	11.6	22.8
30	6	1.8	6.4	11.4	24.2	0.8	5.6	11.2	23.8
60	3	2.0	5.6	11.4	21.6	0.8	5.2	10.0	20.4
60	6	1.2	5.4	11.0	20.0	1.2	5.4	10.8	20.0
		Empirical LR power of c				Empirical LR power of γ			
Treatment effect									
30	3	6.6	11.0	14.8	22.6	1.2	6.8	12.4	22.4
30	6	2.8	7.4	11.0	18.0	1.2	5.8	11.0	18.4
60	3	3.0	6.8	11.0	18.2	1.2	5.4	11.2	21.8
60	6	1.0	4.6	10.8	20.0	1.2	5.0	9.0	19.0

Table 7. Effects of the model misspecification on the significance results (a = 0, ρ = 0.0, c = 0.3).

		Type I error							
		Empirical LR power of ρ				Empirical LR power of β			
N	T	1%	5%	10%	20%	1%	5%	10%	20%
Autocorrelation effect									
30	3	5.8	9.6	13.8	22.6	1.0	6.0	11.0	21.4
30	6	2.0	7.4	13.8	23.2	0.6	6.6	14.0	23.6
60	3	3.2	7.2	12.4	22.2	0.4	5.4	12.8	21.4
60	6	1.4	5.6	12.0	21.6	1.2	5.6	12.0	21.2
		Empirical LR power of c				Empirical LR power of γ			
Treatment effect									
30	3	16.2	25.4	41.8	57.0	8.8	22.0	32.0	49.6
30	6	24.2	50.0	63.0	74.0	19.0	43.8	56.4	67.6
60	3	24.4	50.0	63.0	73.8	16.0	36.4	49.8	63.2
60	6	59.0	82.2	90.0	95.4	48.0	73.6	83.8	92.2

Table 8. Effects of the model misspecification on the significance results (a = 0, ρ = 0.2, c = 0.0).

		Type I error							
		Empirical LR power of ρ				Empirical LR power of β			
N	T	1%	5%	10%	20%	1%	5%	10%	20%
Autocorrelation effect									
30	3	13.6	29.8	42.2	57.2	13.4	29.4	43.2	60.6
30	6	40.4	62.0	74.4	83.0	40.0	61.8	74.6	83.6
60	3	29.6	53.8	67.4	80.6	31.0	56.4	70.4	81.4
60	6	76.8	91.8	96.0	98.0	77.0	91.8	96.0	98.0
		Empirical LR power of c				Empirical LR power of γ			
Treatment effect									
30	3	1.8	6.0	11.2	19.0	1.6	5.4	10.6	22.2
30	6	1.2	5.4	10.8	21.0	1.0	5.6	11.0	20.0
60	3	1.0	5.2	10.0	20.2	0.6	5.8	9.8	20.8
60	6	1.4	5.4	10.4	19.4	1.4	5.2	9.2	19.4

Table 9. Effects of the model misspecification on the significance results (a = 0, ρ = 0.2, c = 0.3).[a]

		Type I error							
		Empirical LR power of ρ				Empirical LR power of β			
N	T	1%	5%	10%	20%	1%	5%	10%	20%
Autocorrelation effect									
30	3	10.2	25.6	38.8	53.6	12.2	27.4	39.2	54.0
30	6	36.2	63.4	73.6	83.0	36.4	63.2	73.6	82.8
60	3	29.2	57.0	66.6	77.8	30.4	58.8	67.4	78.8
60	6	75.4	91.8	95.2	98.4	75.6	92.0	95.2	98.4
		Empirical LR power of c				Empirical LR power of γ			
MLRM									
30	3	10.2	23.6	33.8	47.8	6.0	18.6	27.4	44.0
30	6	17.4	36.8	48.0	59.6	13.0	28.8	39.4	55.6
60	3	18.2	38.6	52.2	65.4	11.6	24.8	35.2	48.0
60	6	39.8	67.4	77.0	87.4	28.2	53.8	68.6	79.8

[a] Longitudinal binary data analysis.

the drug group. Parameters are selected from a combination of the following:

$$a = 0.0,\ c = 0.0,\text{ or } 0.3,\ \rho = 0.0,\text{ or } 0.2,$$
$$N = 30,\text{ or } 60,\quad\text{and } T = 3,\text{ or } 6.$$

Generated data are fitted to the MLRM:

$$\text{logit}\,(p_{it}) = \alpha + \beta\, y_{i,t-1} + \gamma\, \text{TRT}_i.$$

Results of the simulation experiments are: two models are almost equally fitted even though data are generated by the ZLS model, more accurate Type I error rates are achieved in the misspecified MLRM, especially for either smaller sample sizes or smaller tail probabilities, power of testing the autocorrelation (ρ for ZLS, β for MLRM) is similar. However, the power of

testing the treatment effect is a bit less powerful under the misspecified model (MLRM) although it is ignorable for moderately large sample sizes. (For details, see Tables 2–5 for goodness of fit results, and Tables 6–9 for hypothesis testing results.)

Conclusions

With respect to the comparison between the MLRM and ZLS model, the five features (generalizability, interpretability, dealing with incomplete data, software availability, and computability) are considered. The MLRM is preferable in terms of the generalizability and software availability. The effect of model misspecification from ZLS model is ignorable both for conservativeness and for power of the test.

Future research areas are multiple. First, one must explore a more effective model whose characteristics might be interpretability, generalizability, and more fit. Second, one needs to develop a methodology to allow a study with information missing by design and seek for the relative efficiency. The third may be the development of a model that allows a variable with multiple responses. Finally, a more efficient algorithm for statistical inference and its related computer softwares should be developed after performing more extensive comparative studies among the previous models.

REFERENCES

1. Cook, N. R., and Ware, J. H. Design and analysis methods for longitudinal research. Annu. Rev. Public Health 4: 1–24 (1983).
2. Grizzle, J. E., Starmer, C. F., and Koch, G. G. Analysis of categorical data by linear models. Biometrics 25: 357–382.
3. Koch, G. G., Landis, J. R., Freeman, D. A., Freeman, J. L., and Lehnen, R. A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics 33: 133–158 (1977).
4. Liang, K.-Y., and Zeger, S. L. Longitudinal data analysis using generalized linear models. Biometrika 73: 13–22 (1986).
5. McCullagh, P., and Nelder, J. A. Generalized Linear Models. Chapman and Hall, London, 1983.
6. Muentz, L., and Rubinstein, L. V. Markov models for covariates dependence of binary sequences. Biometrika 41: 91–101 (1985).
7. Zeger, S. L., Liang, K-Y., and Self, S. G. The analysis of binary longitudinal data with time-independent covariates. Biometrika 72: 31–38 (1985).
8. Origasa, H. Statistical methods for the analysis of longitudinal data with binary responses. Mimeo Series No. 1853T. Institute of Statistics, University of North Carolina, Chapel Hill, NC, 1988.
9. Liang, K-Y. Odds ratio inference with dependent data. Biometrika 72: 678–682 (1985).
10. Lachin, J. M., and Wei, L. J. Estimators and tests in the analysis of multiple non-independent 2×2 tables with partially missing observations. Biometrics 44: 513–528 (1988).
11. Cochran, W. G. Some methods for strengthening the common chi-square tests. Biometrics 10: 417–451 (1954).
12. Grizzle, J. E., and Allen, D. M. Analysis of growth and dose response curves. Biometrics 25: 357–382 (1969).
13. Dahlquist, G., and Bjorek, A. Numerical methods. Prentice-Hall, Englewood Cliffs, NJ, 1974.